

Introducción a la Estadística

1 INTRODUCCIÓN

Estadística, rama de las matemáticas que se ocupa de reunir, organizar y analizar datos numéricos y que ayuda a resolver problemas como el diseño de experimentos y la toma de decisiones.

2 HISTORIA

Desde los comienzos de la civilización han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o cosas. Hacia el año 3000 a.C. los babilonios usaban pequeñas tablillas de arcilla para recopilar datos sobre la producción agrícola y sobre los géneros vendidos o cambiados mediante trueque. En el siglo XXXI a.C., mucho antes de construir las pirámides, los egipcios analizaban los datos de la población y la renta del país. Los libros bíblicos de Números y Crónicas incluyen, en algunas partes, trabajos de estadística. El primero contiene dos censos de la población de Israel y el segundo describe el bienestar material de las diversas tribus judías. En China existían registros numéricos similares con anterioridad al año 2000 a.C. Los griegos clásicos realizaban censos cuya información se utilizaba hacia el 594 a.C. para cobrar impuestos.

El Imperio romano fue el primer gobierno que recopiló una gran cantidad de datos sobre la población, superficie y renta de todos los territorios bajo su control. Durante la edad media sólo se realizaron algunos censos exhaustivos en Europa. Los reyes carolingios Pipino el Breve y Carlomagno ordenaron hacer estudios minuciosos de las propiedades de la Iglesia en los años 758 y 762 respectivamente. Después de la conquista normanda de Inglaterra en 1066, el rey Guillermo I de Inglaterra encargó la realización de un censo. La información obtenida con este censo, llevado a cabo en 1086, se recoge en el *Domesday Book*. El registro de nacimientos y defunciones comenzó en Inglaterra a principios del siglo XVI, y en 1662 apareció el primer estudio estadístico notable de población, titulado *Observations on the London Bills of Mortality* (*Comentarios sobre las partidas de defunción en Londres*). Un estudio similar sobre la tasa de mortalidad en la ciudad de Breslau, en Alemania, realizado en 1691, fue utilizado por el astrónomo inglés Edmund Halley como base para la primera tabla de mortalidad. En el siglo XIX, con la generalización del método científico para estudiar todos los fenómenos de las ciencias naturales y sociales, los investigadores aceptaron la necesidad de reducir la información a valores numéricos para evitar la ambigüedad de las descripciones verbales.

En nuestros días, la estadística se ha convertido en un método efectivo para describir con exactitud los valores de datos económicos, políticos, sociales, psicológicos, biológicos o físicos, y sirve como herramienta para relacionar y analizar dichos datos. El trabajo del experto estadístico no consiste ya sólo en reunir y tabular los datos, sino sobre todo en el proceso de "interpretación" de esa información. El desarrollo de la teoría de la probabilidad ha aumentado el alcance de las aplicacio-

nes de la estadística. Muchos conjuntos de datos se pueden aproximar, con gran exactitud, utilizando determinadas distribuciones probabilísticas; los resultados de éstas se pueden utilizar para analizar datos estadísticos. La probabilidad es útil para comprobar la fiabilidad de las inferencias estadísticas y para predecir el tipo y la cantidad de datos necesarios en un determinado estudio estadístico.

3 MÉTODOS ESTADÍSTICOS

El método estadístico conlleva una serie de etapas básicas que se describen a continuación:

1. Planteamiento del problema

En primer lugar, debe definirse claramente la población que se desea investigar, sin ambigüedades, así como las variables que debemos observar y cómo se van a medir.

Esta primera fase es fundamental, puesto que las conclusiones que se obtengan sólo podrán aplicarse a la población definida, dependiendo su validez de las variables medidas.

2. Construcción de un modelo estadístico

Deberemos determinar un modelo matemático que se aproxime a la realidad objeto del estudio, de manera que se tengan en cuenta las variables observables más importantes y englobe en la parte aleatoria los efectos de las variables impredecibles.

3. Recogida y presentación de los datos

La recogida de datos se realiza mediante una técnica llamada muestreo. No es necesario analizar el 100% de una población para sacar conclusiones sobre una variable.

Los métodos empleados principalmente para la obtención de los datos son:

- La realización de un experimento repetidas veces. Ej: control de calidad de las aguas.
- Medidas sucesivas. Ej: La precisión de una máquina.
- Cuestionarios a contestar por los individuos. Ej: Análisis de lectura de una población.

Una vez recogidos los datos, se organizan de forma que resulten fácilmente legibles. Para ello, se agrupan en fichas, tablas o cuadros en los que aparecen los posibles resultados de los experimentos y un recuento de las veces que ha ocurrido cada uno de ellos. También se utilizan distintos gráficos para resumir e interpretar fácilmente la información recogida.

4. Depuración de los datos

Tras la recogida de los datos, el estadístico debe proceder a su depuración para detectar los posibles errores.

Los errores pueden ser de varios tipos:

- Error de medida. Va implícito y es inseparable a toda variable cuantitativa, y es debido a los diferentes estados de ánimo de la persona que realiza la medida, a su agudeza visual, a su velocidad de reflejos, a su cansancio, a su habilidad, etc.
- Error de muestreo. Es debido a la falta de uniformidad en las causas que intervienen en la expresión del valor de la variable.
- Error experimental. Es debido a la falta de uniformidad en las condiciones de las experiencias.

La diferencia sistemática entre los resultados obtenidos con la muestra y los datos correctos de toda la población se llama **sesgo**. Para producir datos precisos debemos tomar precauciones para eliminar el sesgo.

Una regla empírica ampliamente contrastada (Huber 1984) es esperar entre un 2 y un 5% de observaciones con errores de medición transcripción, etc.

5. Estimación de los parámetros

Los modelos estadísticos dependen de ciertas constantes, llamadas *parámetros estadísticos*. Utilizando la información de la muestra podremos estimar el valor o los valores de ciertos parámetros, así como el posible error de estimación.

6. Crítica y validación del modelo

En esta fase se comprueba que la información que nos suministra el modelo estadístico y la información empírica es compatible y podemos aceptar el modelo como correcto. En caso contrario deberemos volver al paso 2 para reformular el modelo.

Según los pasos descritos debemos distinguir dos aspectos de la estadística, la llamada **estadística descriptiva**, que recoge, organiza y obtiene los parámetros de una serie de datos, y la llamada **estadística inferencial**, que describe, predice, compara, y generaliza los resultados obtenidos de una muestra a toda la población.

4 LENGUAJE ESTADÍSTICO

El primer campo de actuación de la estadística, como se ha visto, es la demografía. De esta ciencia ha tomado la nomenclatura (población, individuo...).

Se llama **población** al conjunto de todos los elementos cuyo conocimiento interesa. Cada uno de esos elementos es un **individuo**. Si se está estudiando el resultado de ciertos experimentos químicos, cada uno de esos experimentos será un individuo estadístico y el conjunto de todos los posibles experimentos en esas condiciones será la población.

Una **muestra** es el subconjunto de individuos de una población sobre los que se realiza el estudio para trasponer las conclusiones a toda la población.

Cada *individuo* puede ser descrito mediante uno o varios caracteres. Por ejemplo, si los individuos son personas, el sexo, el estado civil, el número de hermanos o su estatura son caracteres. Y si el individuo es una reacción química, el tiempo de reacción, la cantidad de producto obtenido o si éste es ácido o básico serán posibles caracteres que pueden analizarse.

Un **carácter** puede ser **cuantitativo** si es medible numéricamente o **cualitativo** si no admite medición numérica. El número de hermanos y la estatura son caracteres *cuantitativos* mientras que el sexo y el estado civil son caracteres *cualitativos*.

Los distintos valores que puede tomar un carácter cuantitativo configuran una **variable estadística**. La variable estatura, en cierta población estadística, toma valores en el intervalo 147-205; y la variable número de hermanos toma los valores 0, 1, 2, 3, 4, 5, 6, 7 y 8. Una variable estadística como esta última es **discreta**, ya que sólo admite valores aislados. Una variable estadística es **continua** si admite todos los valores de un intervalo, como ocurre con la estatura.

Frecuentemente la información sobre cualquier variable estadística que se pretende estudiar viene dada en forma de tabla. Al realizar el recuento y la organización de los datos obtenidos aparecen los siguientes conceptos:

SF	SF	SF	Modalidad, x_i	Recuento, n_i
IS	SF	IS	IS	7
BI	IS	SF	SF	11
NT	BI	BI	BI	4
SF	NT	NT	NT	5
SB	SB	NT	SB	3
SF	NT	IS		30
IS	IS	SF		
SF	SF	SB		
BI	IS	SF		

Recuento de notas de una clase de Bachillerato

- **Frecuencia absoluta:** es el número de individuos que presentan una modalidad o valor. Se representa por n_i .
- **Frecuencia relativa:** es el cociente entre el número de individuos que presentan una modalidad o valor y el número total de individuos de la población o muestra sobre la que se está realizando el estudio. Se representa por f_i

$$f_i = \frac{n_i}{N}$$

(dónde N es el número de individuos.

- **Porcentaje:** es la frecuencia relativa multiplicada por 100.
- **Frecuencia relativa acumulada:** es la suma de las frecuencias absolutas, una vez ordenados los valores, hasta el que ocupa el lugar i. Se representa por N_i .

$$N_i = \sum_{k=1}^i n_k$$

- **Frecuencia relativa acumulada:** es la suma de las frecuencias relativas, una vez ordenados los valores, hasta la que ocupa el lugar i. Se representa por F_i .

$$F_i = \sum_{k=1}^i f_k$$

5 ENCUESTAS Y SONDEOS

El auge de los sondeos y las encuestas comenzó en 1936 en Estados Unidos. Con motivo de las elecciones presidenciales, una revista se lanzó a realizar una macro-encuesta. Se registraron más de cuatro millones de entrevistas. Se anticipó el ganador, pero las previsiones no se cumplieron.

Fue ese mismo año cuando **George Gallup** efectuó una encuesta a 4500 personas y a partir de los datos hizo una previsión del ganador y acertó, con un pequeño porcentaje de error, el número de votos. A partir de esa fecha, las encuestas por sondeo se empiezan a aplicar en numerosos campos: encuestas de opinión, socioeconómicas, control de producción, estadística media, etc.

- **Sondeo:** es un método de investigación destinado a obtener información de un grupo de individuos previamente seleccionados.
- **Encuesta:** es una técnica que nos va a permitir recoger la información necesaria para un posterior estudio. Esta se puede realizar por observación o, simplemente preguntando a los individuos, y se puede realizar sobre el total de la población o sobre una muestra representativa.
- **Cuestionario:** es el medio de comunicación entre el que solicita los datos y el encuestado, es decir, el instrumento fundamental para la obtención de los datos. Es el documento de trabajo, un conjunto de preguntas en el que se recoge toda la información para ser tratada posteriormente.

Un buen cuestionario debe favorecer la recogida de la información necesaria y facilitar el estudio de los resultados.